ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертацию ГОРШЕНИНА Андрея Константиновича на тему «Полупараметрические методы анализа неоднородных данных и их применение в задачах математического моделирования», представленную на соискание ученой степени доктора физико-математических наук по специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ

Актуальность темы исследования. Современное развитие наук часто опирается на проведение объемных вычислений над большим количеством экспериментальных данных. Несомненный интерес представляют такие наборы данных, которые не укладываются в те классические распределения вероятностей, которые приводятся в руководствах по теории вероятностей и математической статистике. Среди причин такого расхождения между теорией и практикой находится и наличие скрытых, случайных факторов, которые делают изучаемые объекты неоднородными. В связи с этим актуальными являются проблемы выявления новых классов математических моделей в виде параметрических семейств распределений вероятностей, а также разработка и компьютерная реализация алгоритмов идентификации этих распределений.

В диссертации А.К. Горшенина развиваются, во-первых, методы анализа выборок случайного объема, восходящие к классическим работам А.Н. Колмогорова, Б.В. Гнеденко, В.М. Круглова, В.Ю. Королева. Во-вторых, большое внимание уделяется проблематике конечных смесей распределений, включая вопросы оценки параметров, проверки гипотез и интерпретации результатов в динамических задачах. Рассматриваются конечные смеси распределений, отличных от нормальных (гауссовских). Также в диссертации совершенствуется совместное использование методов математической статистики и современных методов машинного обучения и нейронных сетей. Сочетание указанных параметрических и непараметрических методов названо в диссертации *полупараметрическим методом анализа* неоднородных данных.

Перспективность предложенного в диссертации метода подтверждается несколькими примерами его успешного применения для решения задач анализа данных в геологии тел Солнечной системы, в физике турбулентной плазмы, в метеорологии и океанологии.

Характеристика работы и результатов. В целом диссертация оформлена в соответствии с требованиями ВАК.

Во **введении** мотивируется актуальность проблематики диссертационного исследования, приводится обзор монографической и журнальной литературы по данному направлению, перечисляются цели, задачи, методы исследования, сведения о апробации и опубликованности, основные результаты диссертации. Если какой-то параграф содержит новые результаты, тут же приводятся ссылки на авторские публикации этих результатов.

Три первые главы носят теоретический характер. Целью **первой** главы является установление вида распределений некоторых часто встречающихся функций от выборок для выборок случайного объема. В разделе 1.1 напоминается понятие смеси распределений. Также приводится определение динамической компоненты и диффузионной компоненты

дисперсии конечной смеси нормальных законов. Анализ эволюции динамической и диффузной компонент будет в прикладных главах одним из ключевых методов анализа реальных экспериментальных данных. В разделе 1.2 рассматриваются свойства обобщенного отрицательного биномиального распределения. Данное семейство распределений неотрицательной целочисленной случайной величины $N_{r,\gamma,\mu}$ представляет собой смесь распределений Пуассона. При этом параметр распределения Пуассона распределен по обобщенному гамма-распределению (введенному E.W. Stacy в 1962 г.). В итоге возникает трехпараметрическое семейство распределений. Как известно, большое количество параметров позволяет сглаживать более разнообразные выборки. Утверждение 1.2 содержит явные выражения для математического ожидания и дисперсии величины $N_{r,\gamma,\mu}$. В разделе 1.3 доказываются новые утверждения о предельном распределении порядковых статистик (включая максимальный член вариационного ряда) и выборочных квантилей для случайного числа независимых одинаково распределенных случайных величин в предположении об обобщенном отрицательном биномиальном распределении. Вычисляются некоторые начальные моменты предельного распределения для максимума. В разделе 1.4 доказана центральная предельная теорема в схеме серий при случайном числе слагаемых в серии. Введено «случайное условие Линдеберга», обобщающее классическое условие Линдеберга в центральной предельной теореме на случайное число слагаемых. Таким образом, теоретическая и практическая ценность этой главы состоит в обосновании выбора найденных семейств распределений для анализа реальных экспериментальных данных, полученных в подходящих условиях.

Во **второй главе** рассматриваются конечные смеси нормальных законов и конечные смеси гамма-распределений. В разделе 2.1 собраны сведения о классическом, так называемом, EM-алгоритме для разделения смесей, а также о его медианной и стохастической модификациях. Здесь излагаются, в основном, известные в науке факты, хотя есть и новые итерационные формулы (2.7) и (2.8) оценок параметров гамма-распределения, полученные диссертантом. Также здесь вводится идея представить аппроксимации решения уравнения Ланжевена в виде конечных смесей нормальных законов. В разделе 2.2 собраны статистические критерии дли проверки гипотезы о числе компонент, полученные ранее диссертантом. В разделах 2.3–2.5 находятся оценки в вероятностной метрике Леви для конечных смесей нормальных законов с различным числом компонент. Сравниваются расстояния между смесями распределений и расстояния между смешивающими распределениями, что приводит к выводу об устойчивости смесей, т.е. «непрерывности» операции смешивания. Здесь тоже много новых частных результатов. В разделе 2.6 решается часто встречающаяся задача интервального оценивания неизвестного математического ожидания E_X случайной величины X при зашумленных наблюдениях и при округлении наблюдений для ближайшего целого: $Y = [X+\varepsilon+\frac{1}{2}]$. Рассматриваются случай смеси нормальных распределений и случай смеси гамма-распределений для случайной величины X.

В **третьей главе** разрабатываются вопросы построения синтетических процедур анализа динамических изменений структуры потока данных. Эти процедуры появились относительно недавно и получили название метода скользящего разделения смесей (СРС-метод). В разделе 3.1 установлены матричные выражения для математического ожидания,

дисперсии, коэффициентов асимметрии и эксцесса для конечной смеси нормальных законов. В разделе 3.2 для независимых случайных величин X и Y с распределениями в виде конечной смеси нормальных распределений решается задача оценки параметров распределения X по наблюдениям «зашумленного сигнала» Y. Это приводит к необходимости решения переопределенных систем линейных уравнений по методу наименьших квадратов. Показано, что анализ динамической структуры потока данных позволяет также решать известную в статистике случайных процессов *задачу о разладке*. Раздел 3.3 содержит пригодные для компьютерной реализации алгоритмы, позволяющие автоматизировать процесс выделения связанных компонент в СРС-методе. В разделе 3.4 метод определения момента разладки переносится на анализ выделенной с помощью СРС-метода динамической компоненты дисперсии конечной смеси нормальных распределений. В разделе 3.5 предлагается улучшать точность работы СРС-метода с помощью добавления аддитивной «шумовой» компоненты.

Главы с **четвертой** по **шестую** демонстрируют применение предлагаемых полупараметрических методов анализа неоднородных данных для решения сложных задач в геологии, физике плазмы, метеорологии и океанологии, их суммарный объем более 50 % основного текста диссертации. В **четвертой** главе анализируются данные из открытых источниках о распределении размеров пылевых частиц лунного реголита. Здесь продолжаются линия построения статистической модели, восходящая к основополагающим работам А.Н. Колмогорова. Методом минимума расстояния хи-квадрат оцениваются параметры конечной смеси логнормальных распределений с использованием бутстреп-процедур. Это позволяет провести кластерный анализ найденных параметров смесей в пространстве параметров, что в дальнейшем позволит специалистам сопоставить выделенные кластеры с физико-химическими условиями добычи образцов. В **пятой** главе метод анализа данных на основе конечных смесей соединяется с актуальными методами нейронных сетей для прогнозирования. В роли объекта для демонстрации разработанных методов выбраны спектры низкочастотных флуктуаций плазмы, переходные процессы электронно-циклотронного резонансного нагрева плазмы. Удалось осуществить идентификацию амплитудного спектра, выявить повторяемость стохастических процессов, определить и прогнозировать некоторые важные показатели функционирования плазмы. В **шестой** главе развитые в первых трех главах вероятностные модели применяются для исследования метеорологичесих и океанологических данных. Здесь находит естественное применение теория экстремальных наблюдений. Отметим неклассический подход к подгонке отрицательного биномиального распределения к данным на основе минимизации расстояний в функциональных пространствах (вместо метода моментов или метода максимального правдоподобия). Для прогнозирования уровня осадков предлагаются высокоэффективные нейросетевые методы для дискретезированных данных. Предлагается метод определения (классификации) «экстремальности» данного уровня осадков по сравнению с наблюдениями в регионе. Разработанный во второй главе метод оценок коэффициентов в уравнении Ланжевена здесь применяется для моделирования турбулентных потоков тепла между океаном и атмосферой.

В **седьмой** главе приведены сведения о разработанном комплексе программ.

В **Заключении** традиционно формулируются итоги исследования и отмечаются перспективы их дальнейшего использования.

Автореферат верно отражает содержание диссертации.

Достоверность и новизна положений, выводов и рекомендаций. В диссертационной работе строго используются методы теории вероятностей, математической статистики, функционального анализа, линейной алгебры, методов оптимизации, алгоритмы машинного обучения и нейронные сети. Этим обеспечивается достоверность полученных выводов и рекомендаций.

В диссертации получены следующие *основные* новые результаты:

– Смешанные вероятностные модели для выборок со случайным объемом для сумм, максимумов и прореживания по Реньи (Глава 1).

– Доказательства устойчивости в метрике Леви дисперсионно-сдвиговых и конечных сдвиговых смесей нормальных распределений относительно возмущений параметров смешивающего распределения и метода анализа зашумленных и округленных наблюдений (Глава 2).

– Статистическая методология построения моделей сгруппированных скрытых наблюдений при заданных характерных точках их эмпирической функции распределения (Глава 4).

– Комплекс математических и компьютерных методов и алгоритмов статистической идентификации связанных компонент смесей, идентификации и классификации экстремальных наблюдений на основе обобщенных отрицательных биномиальных распределений числа наблюдений и обобщенных гамма-моделей для данных (Главы 3, 5, 6, 7).

Материалы диссертации опубликованы в 82 печатных работах, из которых 31 в журналах из перечня ВАК и 51 статья в Scopus и Web of Science, получены 39 свидетельств о государственной регестрации программ для ЭВМ. Результаты прошли апробацию на международных и российских научных конференциях и семинарах по тематике исследования и внедрены в нескольких научных центрах РФ.

Замечания по работе. По тексту диссертации имеются следующие замечания.

1. При чтении начала раздела 2.6.2 может создаться впечатление, что и случайная величина X и случайные величины ε_j имеют одинаковое распределение с одним и тем же набором параметров $\mathbf{a}$, $\boldsymbol{\sigma}$ и $\mathbf{p}$, см. формулу (2.60) и ссылку на формулу (1.7) в формулировке теоремы 2.13. Но тогда границы доверительных интервалов в задаче про округленные зашумленные наблюдения содержат параметры распределения случайной величины X. Может возникнуть вопрос: если эти параметры неизвестны, то как воспользоваться формулами для доверительных границ? Если параметры известны, то почему бы не воспользоваться явными формулами для математического ожидания? Аналогичное замечание можено сделать про распределения X и ε_j в разделе 2.6.3. Заметим кстати, что не следовало бы использовать один символ j для нумерации наблюдений в обозначении ε_j и для нумерации компонент смеси в формулах (2.60), (2.67).

2. В формуле (2.61) в показателе экспоненты пропущено слагаемое $2\pi n a_j i$, в строках 6 и 7 сверху на стр. 89 в сумме $\sum_{j=1}^{k}$ следует добавить сомножитель $\sin 2\pi n a_j$.

3. На стр. 111 упоминается «метод тестирования однородности на основе критериев Колмогорова–Смирнова и Колмогорова». Далее, на стр. 112 снова упоминается «критерий однородности Колмогорова». В математической статистике Колмогорову принадлежит *критерий согласия*, а критерий однородности для двух выборок называют критерием *Смирнова*. Это два метода для решения разных статистических задач. Как их можно сравнивать и делать вывод, что «подход на основе критерия однородности Колмогорова–Смирнова предпочтительнее» (с. 112)?

4. В разделе 4.3 приводится та формула статистики хи-квадрат (4.5), которая сравнивает относительные частоты с гипотетическими вероятностями. Однако в ней отсутствует множитель вида объема выборки. Отсутствие этого множителя не влияет на процедуру оценки параметров по методу минимума расстояния хи-квадрат, но из текста не ясно, как тогда вычислялись p-значения? При использовании оценок параметров в критерии согласия хи-квадрат число степеней свободы при вычислении p-значения обычно уменьшают на число параметров (теорема Пирсона–Фишера). Это в диссертации не сделано и отсутствуют комментарии по выбору числа степеней свободы для статистики критерия. Более того, как известно, при оценивании параметров не через группированную выборку распределение «статистики хи-квадрат» может отличаться от распределения хи-квадрат Пирсона.

5. К сожалению, автору, глубоко погруженному в тематику своего исследования, при написании текста диссертации на основании впечатляющего числа публикаций легко пропустить некоторые моменты, которые становятся заметны при чтении свежим взглядом. Например, в разделе 3.4 речь ведется о «динамической компоненте волатильности», хотя понятие *волатильности* ранее введено не было и в разделе 1.1 были определены динамическая и диффузионная компоненты *дисперсии конечной смеси*. Название раздела 3.2 обещает «выделение сигнала», но речь в разделе идет об оценке параметров распределения сигнала, а не об удалении шума из наблюдения. На стр. 69 утверждается, что «левое неравенство в формуле (2.18) означает... предположение о конечности дисперсий». Но это неравенство есть $0 < \sigma$, что не исключает случая $\sigma = \infty$. На стр. 73 утверждается, что левое неравенство в (2.33) означает конечность математических ожиданий. Но ведь, не оговаривалось явно, что $a_0 \neq -\infty$? В разделе 3.4 не определяется понятие «события», которые тем не менее надо детектировать.

6. Есть незначительные опечатки, например: в заголовке раздела 5.2.3. надо «анализа» вместо «анализ», на стр. 62 в 6–7 строках пропущено слово «задачи», на стр. 107 в 7й строке снизу должно быть «вектором» вместо «вектор». Общее количество таких огрехов пренебрежимо мало.

Заключение о соответствии диссертации критериям. Сделанные выше замечания не влияют на общую положительную оценку работы. Представленная диссертация

Горшенина А.К. является научно-квалификационной работой, в которой на основании выполненных автором исследований разработаны теоретические положения, совокупность которых можно квалифицировать как научное достижение. Работа выполнена на высоком профессиональном уровне и имеет практическую ценность. В работе представлены как новые математические модели, так и новые численные методы, разработаны комплексы программ.

С учетом всего вышесказанного, диссертация А.К. Горшенина «Полупараметрические методы анализа неоднородных данных и их применение в задачах математического моделирования» полностью соответствует требованиям ВАК к диссертациям на соискание ученой степени доктора физико-математических наук по специальности 05.13.18 — Математическое моделирование, численные методы и комплексы программ. Считаю, что А.К. Горшенин **заслуживает** присуждения ему ученой степени доктора физико-математических наук по указанной специальности.

Официальный оппонент
профессор кафедры программной инженерии
федерального государственного автономного
образовательного учреждения высшего образования
«Национальный исследовательский Нижегородский
государственный университет им. Н.И. Лобачевского»,
доктор физико-математических наук (специальность
01.01.05 – теория вероятностей и математическая
статистика), доцент по кафедре прикладной
теории вероятностей
Зорин Андрей Владимирович

/А. В. Зорин/
26 апреля 2021 г.

Подпись Зорина А.В.
Заверяю. Ученый секретарь ННГУ
Л.Ю. Черноморская
Тел. 462-30-21

Адрес: 603022, г. Нижний Новгород, пр. Гагарина, 23, Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского»,
кафедра программной инженерии
Телефон: (831) 462-33-68, E-mail: andrei.zorine@itmm.unn.ru